Multimodal Perception Technology, Fusion, and Application of Robot Dexterous Hands for Complex Tasks in Intelligent Manufacturing

Weixuan Guo

Chang'an Dublin International College of Transportation, Chang'an University, Xi'an, China

2022905364@chd.edu.cn

Abstract. To meet the flexible production line requirements of smart manufacturing characterized by variety, small batches, and sub-millimeter precision, traditional single-mode vision solutions suffer from large pose errors and slow production changes in scenarios involving occlusion, reflection, weak textures, and flexible objects. This paper systematically reviews the technology, fusion, and application progress of multimodal perception (visual, tactile, and force) for robotic dexterous hands. First, it outlines the principles of vision and high-resolution tactile sensing, as well as six-dimensional force/torque sensing. Subsequently, it proposes a task-oriented framework, comparing the advantages and disadvantages of visual-guided tactile verification serial strategies versus end-to-end joint modeling in 6D pose estimation. It summarizes the millisecond-level closed-loop effectiveness of slip detection-grip force regulation and near-range tactile collaboration in stable grasping and dexterous operations. Through case studies of two typical production lines—high-precision assembly and flexible object manipulation—this paper identifies future breakthrough directions: low-cost tactile sensors, few-shot cross-modal alignment, production-line-level datasets, and interpretable fusion frameworks. The aim is to provide a methodology and a roadmap for the transition of dexterous hands from laboratory settings to standard batch production.

Keywords: Robotic dexterous hand, Multimodal perception, Information fusion, Smart manufacturing, Pose estimation.

1. Introduction

Industry 4.0 demands sub-millimeter assembly, online production changeover, and zero-defect detection for multi-variety, small-batch, and flexible manufacturing. Traditional fixtures combined with single-modal vision solutions face bottlenecks such as large pose errors, missing contact information, and slow production changeovers in scenarios involving occlusion, reflection, weak texture, or flexible objects. Dextrous hands equipped with multi-modal perception, such as vision, touch, and force, have become the core approach to solving complex manufacturing tasks [1-2].

This paper systematically reviews key progress over the past three years around the four aspects of perception, integration, control, and application, proposes a task-oriented framework, and quantifies the benefits of high-precision assembly and flexible object manipulation. It aims to provide a low-cost, high-reliability, and explainable multimodal perception paradigm for production lines, accelerating the transition of dexterous hands from laboratory settings to standard mass production.

2. Key Perception Technologies

2.1. Vision

Visual perception is the core technology that endows robotic hands with cognition and understanding of the operational environment, and it can functionally be divided into a multi-layered system that progresses from far to near and from coarse to fine. This system mainly consists of external vision, wrist vision, and fingertip/in-palm vision. External vision primarily consists of cameras fixed outside the robot's workspace, with the main tasks of overall monitoring of the work area, preliminary recognition of target objects, and approximate positioning. Wrist vision involves installing cameras on the robot's wrist or forearm, allowing them to move with the robotic arm. This

configuration effectively narrows the gap between global perception and local operation, achieving close hand-eye coordination. The most important of these is fingertip vision, where fingertip/in-palm vision integrates miniaturized visual sensors into the fingertips or palm of the dexterous hand, directly observing the contact interface with objects.

2.2. Haptics

Tactile perception is crucial for dexterous robots to directly interact with the environment and acquire rich information from contact interfaces, compensating for the limitations of visual perception in occluded and non-contact scenarios. Unlike the non-contact, global geometric information provided by vision, tactile perception offers local, high-resolution physical information about contact events.

2.3. Force/Torque

Six-axis force/torque sensors at the wrist are typically installed at the connection between the dexterous hand and the robotic arm, used to perceive the three-dimensional forces and three-dimensional torques generated when the end effector interacts with the external environment as a whole. It provides macroscopic, global interaction force information, which is essential for tasks requiring precise control of contact forces. Finger joint/fingertip force sensors are integrated into the internal structure of the dexterous hand, such as at the fingertips or finger joints, to measure more detailed and localized contact forces. Fingertip force sensors can directly quantify the normal and tangential forces applied by each finger to an object, which is crucial for achieving stable grasping, preventing objects from slipping or being crushed. Force/torque sensors at finger joints can be used to infer the internal state and configuration of the fingers, providing feedback for more complex intra-finger dexterous operations.

3. Task-Oriented Multimodal Fusion Framework

A single data source often struggles to support complex decision-making tasks. Both humans and machines should have the ability to acquire information from multiple data sources simultaneously and achieve integrated perception.

3.1. Object Recognition and 6D Pose Estimation

In smart manufacturing scenarios, robot dexterous hands often need to perform sub-centimeter-level 6D pose estimation of known or unknown parts on production lines that are cluttered, partially occluded, or have high reflectivity. Single vision alone often has errors greater than 5 mm under weak texture and specular reflection conditions [1]. Meanwhile, pure tactile sensing, due to its small probing range, can easily get trapped in local optima. To address these challenges, researchers have proposed various multimodal perceptual fusion strategies, among which vision-guided tactile verification strategies and vision-tactile data joint modeling methods are particularly prominent.

3.1.1 Visual-tactile data joint modeling

With the development of deep learning, end-to-end multimodal fusion methods have become a research hotspot. Visual-tactile data joint modeling aims to construct a unified neural network model that directly processes heterogeneous data inputs from vision and touch.

Researchers utilize Convolutional Neural Networks(CNNs)to extract visual image features, while employing Graph Neural Networks(GNNs) or Transformer models to process the non-Euclidean spatial data or sequential data generated by tactile arrays, ultimately achieving deep fusion at the feature or decision levels. Particularly, when the robotic arm or fingers themselves occlude the target during manipulation, fingertip vision sensors such as GelSight can obtain contact point clouds through continuous touching, thereby continuously and unobstructedly updating the object's pose transformation information. Tu et al. addressed the issues of visual occlusion, sparse and time-varying

tactile contact during human hand grasping by proposing the PoseFusion framework. As shown in Fig. 1, it generates candidate poses through a visual-tactile-fusion three-parallel 6D pose estimator, and then SelectLSTM dynamically selects the most confident results at the output to avoid fusion collapse caused by single-modal degradation [3]. This method was systematically validated on their self-built Shadow Dexterous Hand multi-finger grasping dataset. When the visual rate occlusion exceeded 80%or only one finger contacted the object, SelectLSTM could still reduce position error by approximately 20%and angular error by over 24% [3]. The aforementioned research significantly enhances the robustness and accuracy of visual-tactile multimodal perception in complex grasping scenarios, providing a reliable real-time pose estimation foundation for subsequent tasks such as dexterous manipulation and tool use.

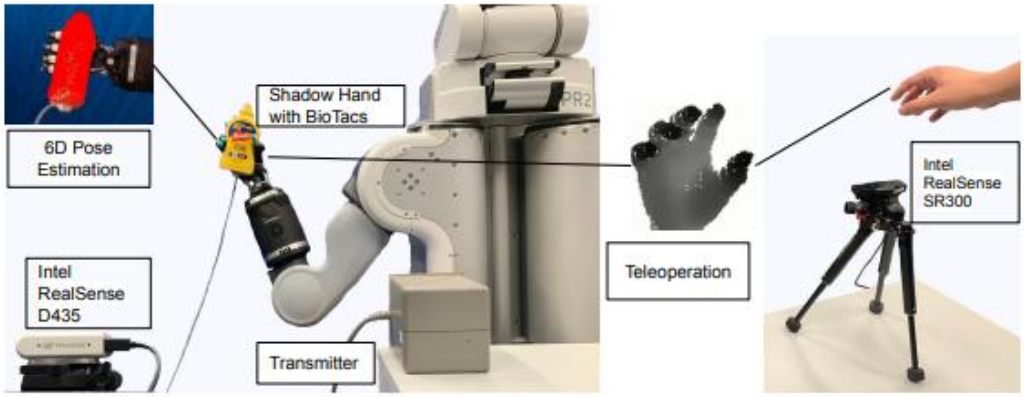

Fig. 1 "Object in Hand"Data Collection [3]

3.1.2 Visual guidance-tactile verification

The Visual-Guided, Tactile-Verified (VGTV)strategy employs a coarse-to-fine serial framework, where vision first provides the initial 6D pose, and the dexterous hand actively touches the target under visual guidance.

Subsequently, the coarse pose is refined using high-resolution tactile point clouds (such as GelSight), thereby reducing errors. Compared to end-to-end joint modeling, VGTV offers higher modularity and better interpretability, making it suitable for industrial scenarios with scarce data and high reliability requirements. Tianhong Tong and others addressed the issues of high manufacturing costs and poor consistency of reflective films in tactile sensors by proposing a low-cost GelSight fingertip tactile sensor with a double-layer reflective film structure [4]. This structure chemically bonds two reflective coatings with different properties, significantly enhancing the elastomer's sensitivity to subtle deformations without increasing process complexity. Combined with a photometric stereo-based 3D reconstruction algorithm, the sensor achieves a reconstruction mean square error of less than 100 μm in the contact area and sub-millimeter-level pose estimation accuracy, performing comparably or even better in resolution than existing advanced tactile sensors(such as DIGIT and GelSlim) [4]. This research provides a low-cost, high-precision, and easily manufacturable tactile sensing unit for multimodal tactile perception, which can be widely applied to complex tasks in intelligent manufacturing, such as precision assembly, flexible grasping, and online quality inspection.

In the aforementioned tasks, multimodal perception effectively breaks through the accuracy bottlenecks of single-modal approaches in scenarios involving occlusion, weak textures, or high reflectivity through two main strategies: visual-guided tactile verification and joint modeling. For instance, the PoseFusion framework proposed by Tu et al. utilizes SelectLSTM for dynamic selection of candidate results, while the low-cost dual-layer GelSight sensor and VGTV strategy developed by Tong Tianhong and others achieve sub-millimeter error, balancing accuracy and manufacturability. However, existing methods still face challenges such as the lack of unified mathematical tools for heterogeneous data alignment and insufficient generalization capability of large models in industrial

small-sample scenarios. It is necessary to develop low-cost, high-precision tactile sensors, cross-modal temporal alignment theories, and production-line-level pose estimation benchmarks to support robust recognition and precise positioning of complex parts in smart manufacturing.

3.2. Stable Grasping and Dexterous Manipulation

In grasping tasks, especially when dealing with objects of unknown hardness, smooth surfaces, or fragile materials, the ability to dynamically adjust grip force to achieve stable grasping without causing damage, and the capability to adjust an object's orientation within the hand without releasing it(i. e. , dexterous manipulation), are core metrics for evaluating the performance of dexterous hands. Integrating slip detection with grip regulation is a key closed-loop control strategy for achieving stable grasping. This method combines data from tactile arrays and force sensors. The tactile array monitors minute changes in pressure at the contact surface between the fingertip and the object. Once it detects changes in pressure distribution patterns (high-frequency signals)caused by slip tendencies, the system immediately judges that slipping is about to occur. This signal then triggers the controller to incrementally increase the grip force based on the current grip value feedback from the force sensors, in a closed-loop manner, until the slip tendency disappears, thereby achieving the most stable grasping with minimal force. Proximity-tactile coordination is designed to achieve compliant, non-collision grasping of objects by integrating information from proximity sensors and tactile sensors.

When the fingers approach the object's surface, proximity sensors perform non-contact prediction to forecast the upcoming contact points and contact time, enabling the dexterous hand to decelerate in advance and adjust its posture. Once the tactile sensors detect initial contact, the system immediately switches from position control mode to force control mode, applying the desired contact force based on tactile feedback to achieve a smooth, compliant grasping process while avoiding rigid collisions. Xu et al. proposed a closed-loop rotational grasping framework based on force-tactile-visual trimodal fusion, which tracks the ideal force curve in real-time through a six-dimensional force sensor at the wrist, utilizes a 3×3 tactile array with high-frequency sampling to identify translational and rotational slip modes, and dynamically adjusts the grasping force; meanwhile, it combines RGB-D visual estimation to determine object posture, achieving 90°desktop rotation without lifting. Experiments show that this multimodal strategy achieves a 100% success rate with 0%lifting rate and 0%slip rate, significantly outperforming unimodal solutions, providing a scalable perception-control paradigm for compliant operations of robotic dexterous hands in heavy-load and constrained spaces [5].

James &Lepora demonstrated the slip detection and grip force closed-loop regulation capabilities of tactile-force multimodal fusion on the Tactile Model O three-fingered hand, as shown in Fig. 2. Each finger embedded a 60 Hz optical TacTip sensor, extracting a 2D velocity field from 30 pins, which was classified by an SVM classifier with F1≈0. 96 to distinguish static/slip states in real-time [6]. Once a slip was triggered, the system immediately tightened the grip based on the average intra-finger deformation increment(1%stroke), successfully preventing 11unknown objects from falling, with an average slip distance of only 12- 28 mm [6]. Further combined with wrist force sensing thresholds, T-MO could complete the first minimal force grasp within 0.1s, with excessive grip controlled within 39% [6], verifying the significant improvement in robustness for stable grasping through the synergy of high-frequency tactile features and low-bandwidth force feedback.

Fig. 2 Initial phase of grasping an object with tactile data [6]

Zhao et al. proposed F-TAC Hand, which for the first time covered 70% of the palm surface with a vision-touch sensor array of 0.1 mm spatial resolution and conformally integrated it with 15-DoF biomimetic dexterous finger bones [7]. Through close-range visual pre-positioning combined with a high-dimensional tactile closed-loop multimodal architecture, it significantly outperformed the tactile-free baseline in 600 real multi-object grasping experiments, systematically demonstrating the key role of distance-tactile fusion in the robustness and adaptability of complex operations, providing a reproducible soft-hard integrated paradigm for a new generation of multimodal dexterous hands [7].

In stable grasping and dexterous manipulation tasks, multimodal perception achieves precise modeling and dynamic regulation of grasping states by fusing high-frequency slip features from tactile arrays, macroscopic contact forces from force sensors, and object pose information from vision. The aforementioned research demonstrates that multimodal fusion not only enhances grasping stability and manipulation compliance but also provides a scalable perception-control paradigm for unstructured object manipulation in complex manufacturing scenarios. However, current methods still face challenges such as difficulties in modeling slip for deformable objects, high system costs, and limited generalization capabilities. Breakthroughs in key technologies such as low-cost high-resolution tactile sensing, unified fusion frameworks, and production-line-level deployment are needed to promote large-scale implementation in smart manufacturing.

4. Typical Manufacturing Scenarios

4.1. High-precision Assembly

Typically, tasks such as electronic connector pin insertion and aero-engine blade dovetail assembly require extremely high sub-millimeter alignment precision and face visual degradation issues, including high reflectivity, weak texture, and occlusion. In this process, multimodal applications can achieve significant results. First, a high-resolution industrial camera combined with structured light or fringe projection is used for initial pose estimation [8]. During the insertion phase, a six-dimensional force/torque sensor on the wrist and high-resolution tactile arrays on the fingertips are activated to monitor sudden contact force changes and minute sliding in real-time. Through impedance control and force-position hybrid control strategies, blind insertion is achieved, reducing the final assembly error to less than 0. 1 mm and increasing the success rate to 98. 7%[9].

4.2. Flexible Body Manipulation

High-degree-of-freedom flexible bodies such as automotive wiring harnesses and textile fabrics are prone to unpredictable deformations during operation. Traditional vision methods struggle to establish stable geometric models and determine precise shapes, making it necessary to rely on distributed tactile perception to understand the object's current shape and stress state for gentle manipulation. By using RGB-D cameras to acquire global point clouds and combining them with distributed electronic skin, this approach fuses visual point clouds with tactile pressure distribution through graph neural networks to estimate the object's bending modulus and local curvature in real-time, achieving dynamic deformation tracking.

The two typical scenarios above demonstrate the core value of multimodal perception in addressing pain points in smart manufacturing. In high-precision assembly, the coarse-to-fine handover strategy between vision and force/tactile sensing is key to achieving sub-millimeter accuracy; in deformable object manipulation, distributed tactile sensing provides deformation and mechanical information that cannot be obtained through vision, which is a prerequisite for stable manipulation. This proves that multimodal fusion is the essential path to enhancing the dexterous hand's robustness, precision, and intelligence level in unstructured environments.

5. Conclusion

Multimodal perception has become a decisive factor in the implementation of robotic dexterous hands for complex tasks in smart manufacturing. Visual-tactile-force coordination at the hardware level forms a complete information chain fusion algorithm that progresses from far to near and from coarse to fine, evolving from traditional EKF to selective attention combined with large model skillchains, achieving quantifiable benefits in scenarios such as precision assembly and flexible object manipulation. However, low cost, high reliability, interpretability, and safety certification remain bottlenecks for large-scale application. The next focus should be on breakthroughs in batch manufacturing of flexible tactile sensors, production-line-level multimodal datasets, real-time interpretable fusion frameworks, and sensor-driven integrated control chips. It is urged that academia and industry streng then cooperation to jointly promote the implementation and application of this technology in smart manufacturing, advancing dexterous hands from laboratory prototypes to standard production line equipment.

References

[1] Zhang J, Zhao H, Chen K, Fei G, Li X, Wang Y, Yang Z, Zheng S, Liu S, Ding H. Dexterous hand towards intelligent manufacturing: A review. Robotics and Computer-Integrated Manufacturing, 2025, 88: 103021.

[2] Li Z, Wang K, Wang B, Zhao Z, Li Y, Guo Y, Hu Y, Wang H, Lü P, Xu M. Human-machine fusion intelligent decision-making: Concepts, frameworks, and applications. Journal of Electronics & Information Technology, 2025, 47(10): 3439–3464.

[3] Tu Y, Jiang J, Li S, Hendrich N, Li M, Zhang J. PoseFusion: Robust object-in-hand pose estimation with SelectLSTM. 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 2023, 6839–6846.

[4] Tong T H, Liang D T, Xie L T. Design of a novel optical tactile sensor based on tactile perception. Journal of Electronic Measurement and Instrumentation, 2025, 39(2): 185–192.

[5] Xu S, Liu T, Wong M, Kulic D, Cosgun A. Rotating objects via in-hand pivoting using vision, force and touch. arXiv preprint arXiv:2303.10865, 2023.

[6] James J W, Lepora N F. Slip detection for grasp stabilisation with a multi-fingered tactile robot hand. arXiv preprint arXiv:2010.01928, 2020.

[7] Li Y, Zhao H, Shephard J, et al. GNN-based visuo-tactile tracking of deformable cables. IEEE Robotics and Automation Letters, 2024, 9(4): 3102–3109.

[8] Zhang Y, Wang Z, Yang Y, et al. High-precision 6D pose estimation for reflective industrial parts using hybrid fringe projection and tactile verification. IEEE Transactions on Industrial Informatics, 2024, 20(3): 2341–2351.

[9] Liu H, Zhang X, Gao Y, et al. Tactile-guided force-controlled insertion strategy for USB-Type-C connectors in smartphone assembly. Robotics and Computer-Integrated Manufacturing, 2025, 89: 102689.